Class Semantic Attribute Perception Guided Zero-Shot Learning

Qin Yue, Junbiao Cui, Jianqing Liang*, Liang Bai*

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, China
993203718@qq.com,{cjb, liangjq, bailiang}@sxu.edu.cn

Abstract

Deep learning has achieved remarkable success in supervised image classification tasks, which relies on a large number of labeled samples for each class. Recently, zero-shot learning has garnered significant attention, which aims to recognize unseen classes using only training samples from seen classes. To bridge the gap between images and classes, class semantic attributes are introduced, making the alignment between image and class semantic attributes critical to zero-shot learning. However, existing methods often struggle to accurately focus on the image regions corresponding to individual class semantic attributes and tend to overlook the relations between different regions of an image, leading to poor alignment. To address these challenges, we propose a class semantic attribute perception guided zero-shot learning method. Specifically, we achieve coarse-grained perception of class semantic attributes across the entire image through contrastive semantic learning. Additionally, we attain fine-grained perception of individual class semantic attributes within image regions via region partitioning-based attribute alignment, which fully considers the relations between different regions of an image. By integrating these two processes into a unified network, we achieve multi-grained class semantic attribute perception, thereby enhancing the alignment between images and class semantic attributes. We validate the effectiveness of the proposed method on zero-shot learning benchmark data sets.

Introduction

Deep learning has made remarkable progress in fully supervised image classification tasks (Russakovsky et al. 2015; He et al. 2016; Dosovitskiy et al. 2021; He et al. 2022). However, labeling samples for every class is impractical, especially as new classes continually emerge. In response, zero-shot learning (ZSL) (Xian et al. 2019a; Xu et al. 2022a; Tang et al. 2022; Zhang et al. 2023; Chen et al. 2022c; Xu et al. 2022b; Narayan et al. 2021; Chen et al. 2021a) has recently gained significant attention. The goal of ZSL (Larochelle, Erhan, and Bengio 2008) is to recognize unseen classes for which no training samples are available. To achieve this, class semantic attributes (Frome et al. 2013; Lampert, Nickisch, and Harmeling 2009; Reed et al. 2016) are typically introduced as a bridge between images and classes. These attributes, shared between seen and unseen classes, enable ZSL methods to leverage the knowledge acquired from seen classes to effectively recognize unseen classes.

In zero-shot image classification, image embeddings are typically extracted by convolutional neural networks such as VGG (Simonyan and Zisserman 2015), GoogleNet (Szegedy et al. 2015), ResNet (He et al. 2016), etc. Meanwhile, class semantic attributes are derived from expert annotations or language models (Xian et al. 2019a; Welinder et al. 2010; Patterson and Hays 2012). These image embeddings are then aligned with their corresponding class semantic attributes. However, due to the significant differences between visual images and class semantic attributes, the primary challenge in zero-shot image classification is effectively aligning images with class semantic attributes.

According to how to achieve alignment between images and class semantic attributes, the existing zero-shot image classification methods can be roughly categorized into two types, i.e., overall embedding-based alignment methods (see Fig. 1a) and part embedding-based alignment methods (see Fig. 1b).

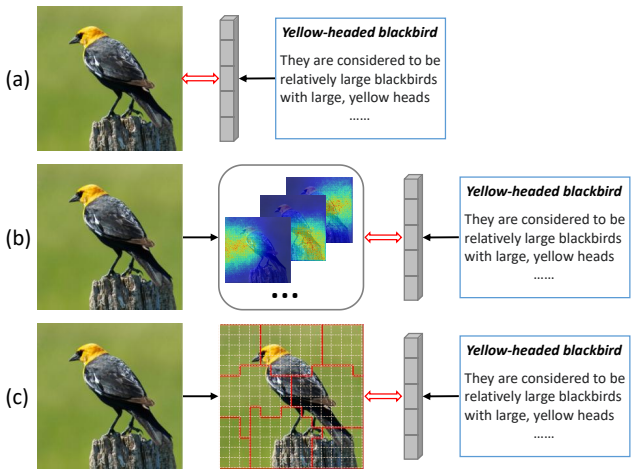


Figure 1: An illustration of different alignment methods. (a) Overall embedding-based alignment method. (b) Part embedding-based alignment method. (c) The proposed method.

*Corresponding authors.

Overall embedding-based alignment methods (see Section Related Works) were the first to be proposed and have garnered significant attention. This kind of method focuses on extracting an overall embedding of an image and aligning it with the corresponding class semantic attributes. In ZSL, the class semantic attributes serve as transferable knowledge shared between seen and unseen classes. However, these methods struggle to effectively achieve the perception of individual class semantic attributes, which limits transferability of knowledge from seen classes to unseen classes.

More recently, part embedding-based alignment methods (see Section Related Works) have been introduced. This kind of method aims to align the part embedding of an image with the corresponding class semantic attributes. However, they still face several limitations. First, these methods often focus on the parts of an image that are too broad, failing to effectively focus on the region corresponding to individual class semantic attributes. Second, they rarely account for the relations between different regions of an image during the alignment process between images and class semantic attributes.

To tackle the above challenges, we propose a class semantic attribute perception guided zero-shot learning method (CSAP-ZSL) (see Fig. 1c). The regions of an image are divided into different clusters. The regions of an image within the same cluster are aligned with the corresponding class semantic attributes, achieving fine-grained perception. In this paper, the proposed CSAP-ZSL incorporates both coarse-grained and fine-grained perception modules. In the coarse-grained perception module, we first learn the overall embedding of an image. Based on contrastive semantic learning, we achieve coarse-grained perception of class semantic attributes across the entire image. In the fine-grained perception module, we split an image into some small regions. Then we explore the relations between different regions of an image through soft region partitioning. Furthermore, we achieve fine-grained perception of individual class semantic attributes within image regions through region partitioning-based attribute alignment. Finally, we integrate these two modules into a united network, enabling multi-grained class attribute perception for ZSL.

To summarize, the main contributions of this paper are as follows:

- We propose a class semantic attribute perception guided ZSL method, incorporating both coarse-grained and fine-grained perception modules.
- In the coarse-grained perception module, we achieve coarse-grained perception of class semantic attributes across the entire image through contrastive semantic learning.
- In the fine-grained perception module, we achieve fine-grained perception of individual class semantic attributes within image regions. We sufficiently consider the relations between the different regions of an image through region partitioning-based attribute alignment.
- Extensive experiments on ZSL benchmark data sets validate the effectiveness of the proposed method compared to state-of-the-art methods.

Related Works

Overall Embedding-based Alignment Method

Overall embedding-based alignment methods can be roughly divided into embedding methods and generative methods. Embedding methods (Romera-Paredes and Torr 2015; Kodirov, Xiang, and Gong 2017; Zhang, Xiang, and Gong 2017; Song et al. 2018; Li et al. 2018, 2020; Zhang, Liang, and Zhao 2022; Zhang et al. 2023) aim to learn a mapping between visual image space and class semantic attributes space on training samples from seen classes. At test phase, the samples are mapped into embedding space and are classified by nearest neighbor search. Because the learned model is used for recognizing unseen classes without any adaptation, the domain shift problem (Fu et al. 2015, 2018) will be caused. Later, a large number of generative methods are proposed (Schönfeld et al. 2019; Xian et al. 2019b; Li et al. 2019; Chen et al. 2021a,b; Guan et al. 2021; Han et al. 2021, 2022; Kong et al. 2022). These methods aim to learn a generator that can generate the image features for unseen classes given class semantic attributes. Then the ZSL problem is converted into a fully supervised learning problem.

However, the above methods are unable to effectively achieve the perception of individual class semantic attributes on an image, which limits transferability of knowledge from seen classes to unseen classes.

Part Embedding-based Alignment Method

Recently, part embedding-based alignment methods (Xie et al. 2019; Zhu et al. 2019; Huynh and Elhamifar 2020; Chen et al. 2022d; Xu et al. 2020; Chen et al. 2022b,a, 2023, 2022c) are proposed. These methods focus on the part of an image and enable it to align with corresponding class semantic attributes. Among them, AREN (Xie et al. 2019) and SGMA (Zhu et al. 2019) attempt to mask or crop the image parts to conduct alignment between image and class semantic attributes. DAZLE (Huynh and Elhamifar 2020) and MSDN (Chen et al. 2022d) adopt the dense attention mechanism to locate the image parts of the class semantic attributes. APN (Xu et al. 2020) designs an attribute prototype network to locate the class semantic attributes in an image. TransZero (Chen et al. 2022b) learns a discriminative image feature by exploiting the Transformer-like mechanism to achieve the guidance of the class semantic attributes. Further, TransZero++ (Chen et al. 2022a) and DUET (Chen et al. 2023) adopt Transformer-like mechanism for image and class semantics, respectively, which achieves the mutual guidance between the image embedding and class semantic attributes. GNDAN (Chen et al. 2022c) learns more discriminative embedding of the image by attention network, which realizes the interactions between visual space and class semantic space.

However, the above methods learn the part embedding is too broad to effectively focus on individual class semantic attributes and ignore the relations between different regions of an image during the process of alignment between images and class semantic attributes.

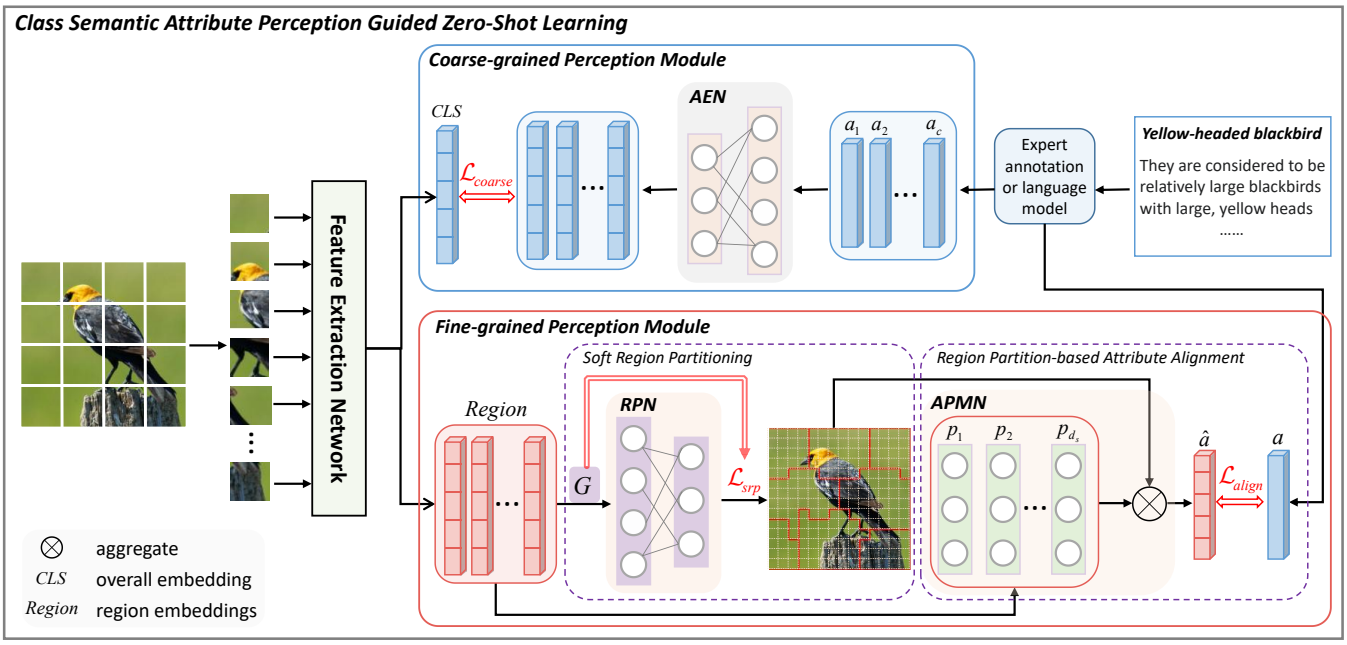


Figure 2: The overview of the proposed CSAP-ZSL. The method includes coarse-grained and fine-grained perception modules. G denotes the similarity graph of regions. AEN denotes attribute embedding network. RPN denotes region partitioning network. APMN denotes attribute prototype matching network.

Problem Formalization

Given an image classification task, let $\mathcal{X}$, $\mathcal{Y}$ and $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$ be the input space, the class label space and the unknown classification function, respectively. Specifically, $\mathcal{X} \subseteq \mathbb{R}^{C\times H\times W}$, where C is the number of channels of an image, H and W is the height and the width of an image, respectively. In ZSL, the class label space consists of two disjoint parts, i.e., $\mathcal{Y} = \mathcal{Y}_s \cup \mathcal{Y}_u$ and $\mathcal{Y}_s \cap \mathcal{Y}_u = \phi$. For the sake of discussion, let $\mathcal{Y}_s = \{1, 2, \cdots, s\}$ and $\mathcal{Y}_u = \{s+1, s+2, \cdots, s+u\}$ be the set of seen classes and unseen classes, respectively. And $c = s+u$ denotes the total number of the classes. In addition, $\forall k \in \mathcal{Y}$, the class semantic attributes vector $\mathbf{a}_k \in \mathbb{R}^{d_s}$ is introduced to establish the bridge between seen classes and unseen classes.

Let $D_{tr} = \{(x_i, y_i)\}_{i=1}^{n}$ and $D_{te} = \{x_j\}_{j=n+1}^{n+m}$ be the training and test data sets, respectively. In ZSL, all the training samples come from seen classes, i.e., $\forall (x_i, y_i) \in D_{tr}$, $y_i = \varphi(x_i) \in \mathcal{Y}_s$. In the test phase, samples can come from the whole class space, i.e., $\forall x_j \in D_{te}$, $\varphi(x_j) \in \mathcal{Y}$.

Methodology

The overview of the proposed method CSAP-ZSL is shown in Fig. 2, which incorporates both fine-grained and coarse-grained perception modules. To more effectively capture the region features of an image, we split an image into some small regions. In the fine-grained perception module, we first explore the relations between different regions of an image through soft region partitioning. Subsequently, we conduct the alignment between the partitioning result and the class semantic attributes, which can realize fine-grained perception of individual class semantic attributes within image regions. Meanwhile, in the coarse-grained perception module, we achieve coarse-grained perception of class semantic attributes across the entire image through contrastive semantic learning.

The total objective function is

$$\mathcal{L} = \mathcal{L}_{fine} + \mathcal{L}_{coarse}, \tag{1}$$

where $\mathcal{L}_{fine}$ and $\mathcal{L}_{coarse}$ denote the fine-grained perception loss and coarse-grained perception loss, respectively. Detailed descriptions for each module of the proposed method are given in the following sections.

Fine-grained Perception Module

This module aims to achieve fine-grained perception of individual class semantic attributes within image regions. This module includes two components: soft region partitioning and region partition-based attribute alignment. The soft region partitioning component sufficiently explores the relations between different regions of an image. Furthermore, the region partition-based attribute alignment component ensures accurate alignment between image regions within the same cluster and the corresponding class semantic attributes.

Given an image $x_i \in \mathbb{R}^{C\times H\times W}$, let $split(x_i) = (x_{i_1}, x_{i_2}, \cdots, x_{i_P})$ be the split result. Specifically, $\forall p = 1, 2, \cdots, P$, $x_{i_p} \in \mathbb{R}^{C\times H'\times W'}$ is the pth image region, where H' and W' are the height and width of the image region, respectively. And $P = \frac{H}{H'} \times \frac{W}{W'}$ is the total number of regions of an image [1].

[1]For simplicity, the divisible case is considered here.

Let $f_e\left(split(x_i);\Theta_1\right) = (\mathbf{h}_{i_0}, \mathbf{h}_{i_1}, \mathbf{h}_{i_2}, \cdots, \mathbf{h}_{i_P})$ be the embedding of the split result output by feature extraction network f_e. Specifically, $\mathbf{h}_{i_0} \in \mathbb{R}^d$ is the overall embedding of image x_i. $\forall p = 1, 2, \cdots, P$, $\mathbf{h}_{i_p} \in \mathbb{R}^d$ is the embedding of the pth region of image x_i. d is the dimension of embedding. Θ_1 is the set of learnable parameters of f_e.

Soft Region Partitioning The goal of soft region partitioning is to explore the relations between different regions of an image, where the regions within the same cluster correspond to individual semantic attributes. To achieve this, we model the soft region partitioning problem as a graph cut problem. For an image x_i, we construct an undirected graph $\mathcal{G}_i = (\mathcal{V}_i, \mathbf{S}_i)$, where $\mathcal{V}_i = \{v_{i_1}, v_{i_2}, \cdots, v_{i_P}\}$ represents the set of vertexes and $\mathbf{S}_i \in \mathbb{R}^{P\times P}$ represents the weights of edges on graph $\mathcal{G}_i$. The vertexes correspond to regions of an image. $s_{i_{pq}}$ denotes the weight of the edge between vertices v_{i_p} and v_{i_q}, calculated by the following formula

$$s_{i_{pq}} = \begin{cases} 1, & \text{if } sim(v_{i_p}, v_{i_q}) > \tau \ and \ q \in radius_r(p) \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where the $radius_r(p)$ denotes the image regions within radius r of pth region, which characterizes the spatial position relation between image regions. τ is hyper-parameter and is set to $1e^{-5}$ in the proposed method. $sim(v_{i_p}, v_{i_q})$ denotes the similarity between the pth region and the qth region and cosine similarity is adopted, i.e.,

$$sim(v_{i_p}, v_{i_q}) = \frac{\mathbf{h}_{i_p}^T \mathbf{h}_{i_q}}{\|\mathbf{h}_{i_p}\|_2 \|\mathbf{h}_{i_q}\|_2}. \tag{3}$$

In addition, we design the region partitioning network (RPN) to learn the graph cut result, and we enable the assigning label to satisfy the probability distribution. Therefore, for each region of an image, we have

$$f_p\left(\mathbf{h}_{i_p};\Theta_2\right) \in \left\{ \mathbf{v} \,\middle|\, \forall k = 1, \cdots, K, v_k \geq 0, \sum_{k=1}^{K} v_k = 1 \right\}, \tag{4}$$

where $\mathbf{h}_{i_p}$ denotes the embedding of the pth region of image x_i. f_p denotes RPN and the Θ_2 is the set of learnable parameters of f_p. K is the number of clusters. Finally, we constrain similar regions to be assigned to the same cluster, which can preserve the similarities between different regions of an image. Therefore, the loss of soft region partitioning can be formalized as

$$\mathcal{L}_{srp} = \frac{1}{nP^2} \sum_{i=1}^{n} \sum_{p,q=1}^{P} \left\| f_p\left(\mathbf{h}_{i_p};\Theta_2\right) - f_p\left(\mathbf{h}_{i_q};\Theta_2\right) \right\|_2^2 s_{i_{pq}}. \tag{5}$$

Region Partition-based Attribute Alignment This component aims to achieve accurate alignment between image regions within the same cluster and the corresponding class semantic attributes. As shown in Fig. 2, we design attribute prototype matching network (APMN) to calculate the response score of the each region of an image for each class semantic attribute. We have

$$f_m\left(\mathbf{h}_{i_p};\mathbf{P}\right) = \mathbf{h}_{i_p}^T \mathbf{P}, \tag{6}$$

where $\mathbf{P} = \left(\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_{d_s}\right) \in \mathbb{R}^{d\times d_s}$ and $\mathbf{p}_j$ denotes the jth class semantic attribute prototype and can be updated in the learning process. Furthermore, we calculate the predicting value of an image for jth class attribute prototype by the following formula

$$\hat{a}_{ij} = \underset{k=1,2,\cdots,K}{max} \sum_{p=1}^{P} m_{i_{pk}} f_m\left(\mathbf{h}_{i_p};\mathbf{P}\right)_j, \forall j = 1, 2, \cdots, d_s, \tag{7}$$

where $m_{i_{pk}} = f_p\left(\mathbf{h}_{i_p};\Theta_2\right)_k$ is the membership of the pth region of image x_i corresponding to the kth cluster. We use the maximum score of regions belonging to the same cluster matching with class semantic prototype $\mathbf{p}_j$ as the predicting value for the jth attribute. Finally, we use the regression loss on the predicting class semantic attributes and true class semantic attributes, i.e.,

$$\mathcal{L}_{align} = \frac{1}{n} \sum_{i=1}^{n} \|\hat{\mathbf{a}}_i - \mathbf{a}_{y_i}\|_2^2, \tag{8}$$

where $\hat{\mathbf{a}}_i$ is the predicting class semantic attributes of image x_i by formula (7). y_i denotes the true class label of image x_i and $\mathbf{a}_{y_i}$ denotes the true class semantic attributes of y_i.

To conclude, the fine-grained perception loss is formalized as

$$\mathcal{L}_{fine} = \alpha\mathcal{L}_{srp} + \beta\mathcal{L}_{align}, \tag{9}$$

where $\alpha > 0$ and $\beta > 0$ are the trade-off parameters.

Coarse-grained Perception Module

The goal of this module aims to achieve coarse-grained perception of class semantic attributes across the entire image through contrastive semantic learning. As shown in Fig. 2, we design an attribute embedding network (AEN) to map the class semantic attributes into the latent space. At the same time, we learn the image embedding by the feature extraction network f_e. Finally, we conduct contrastive semantic loss in the latent space. The coarse-grained perception loss is formalized as

$$\mathcal{L}_{coarse} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{\exp\left(\mathbf{h}_{i_0}^T f_a(\mathbf{a}_{y_i};\Theta_3)\right)}{\sum_{k\in\mathcal{Y}_s} \exp\left(\mathbf{h}_{i_0}^T f_a(\mathbf{a}_k;\Theta_3)\right)}, \tag{10}$$

where $\mathbf{h}_{i_0}$ denotes the embedding of the image x_i obtained by f_e, and f_a denotes AEN. The overall embedding of the image is constrained to align with the corresponding class semantic attributes $\mathbf{a}_{y_i}$ to achieve the coarse-grained perception of class semantic attributes. y_i denotes the true label of the image x_i. Θ_3 is the set of learnable parameters of the network f_a, and $f_a\left(\mathbf{a}_{y_i};\Theta_3\right) \in \mathbb{R}^d$.

Predict

First, the test image $x_j \in D_{te}$ is mapped into the latent space. Then, we search for the nearest class semantic embedding of the test image to assign the class label. Thus, we have

$$\hat{y}_j = \underset{k\in\mathcal{Y}}{arg\,max} \ \ f_e\left(x_j;\Theta_1^*\right)^T f_a\left(\mathbf{a}_k;\Theta_3^*\right) - \gamma\mathbb{I}[k \in \mathcal{Y}_s], \tag{11}$$

where Θ_1^* and Θ_3^* denote the learned parameters. $\mathbb{I}[k \in \mathcal{Y}_s] = 1$ if k corresponds to seen classes and $\mathbb{I}[k \in \mathcal{Y}_s] = 0$ otherwise. $\gamma > 0$ is the calibration factor.

Experiments

Experimental Settings

Data Sets In the experiments, three benchmark ZSL data sets AWA2 (Animals with Attributes2) (Xian et al. 2019a), CUB (Caltech-UCSD Birds-200-2011) (Welinder et al. 2010), and SUN (SUN Attribute) (Patterson and Hays 2012) are used for evaluating the performance of the proposed method. The data sets and data split both follow the literature (Xian et al. 2019a). The detailed information of data sets is summarized in Table 1.

Data Set	d_s	$\lvert\mathcal{Y}\rvert$	$\lvert\mathcal{Y}_s\rvert$	$\lvert\mathcal{Y}_u\rvert$	Total Samples	Training Samples		Test Samples	
						$\mathcal{Y}_s$	$\mathcal{Y}_u$	$\mathcal{Y}_s$	$\mathcal{Y}_u$
AWA2	85	50	40	10	37322	23527	0	5882	7913
CUB	312	200	150	50	11788	7057	0	1764	2967
SUN	102	717	645	72	14340	10320	0	2580	1440

Table 1: The basic information of three zero-shot classification benchmark data sets.

Comparison Methods For overall embedding-based alignment ZSL methods, we select some embedding methods, which include QFSL (Song et al. 2018), LDF (Li et al. 2018), EBSG (Zhang, Liang, and Zhao 2022), DVBE (Min et al. 2020) and mVACA (Zhang et al. 2023). In addition, we select some generative methods, which include CADA-VAE (Schönfeld et al. 2019), f-VAEGAN (Xian et al. 2019b), LisGAN (Li et al. 2019), FREE(Chen et al. 2021a), HSVA (Chen et al. 2021b), BPL (Guan et al. 2021), CE-GZSL (Han et al. 2021), SCE-GZSL (Han et al. 2022) and ICCE(Kong et al. 2022). For part embedding-based alignment ZSL method, we select some non-end-to-end comparison methods, which include DAZLE(Huynh and Elhamifar 2020), TransZero (Chen et al. 2022b), TransZero++ (Chen et al. 2022a), MSDN(Chen et al. 2022d) and GNDAN (Chen et al. 2022c). In addition, we select some end-to-end comparison methods, which include SP-AEN (Chen et al. 2018), AREN (Xie et al. 2019), LFGAA (Liu et al. 2019), SGMA (Zhu et al. 2019), APN (Xu et al. 2020), VSE (Zhu, Wang, and Saligrama 2019) and DEUT (Chen et al. 2023).

Evaluation Protocol In the test phase, we follow the unified evaluation protocols in literature(Xian et al. 2019a). Specifically, for the generalized ZSL task, we report the top-1 accuracy on seen classes (S) and unseen classes (U), as well as their harmonic mean (H = (2 × S × U)/(S + U)).

Implementation Details In the proposed method CSAP-ZSL, we take the vision transformer (Dosovitskiy et al. 2021) pre-trained on ImageNet1K as feature extraction network. In addition, some parameters need to be determined. Specifically, the trade-off parameters α and β both are searched in the set $\{1e^{-3}, 5e^{-4}, 1e^{-4}, 5e^{-5}, 1e^{-5}\}$. The number of clusters in graph cut K is set to 7, 8, and 8 for data sets AWA2, CUB, and SUN, respectively. The radius of the similarity graph of regions is set to 6, 5, and 10 for data sets AWA2, CUB, and SUN, respectively. The calibration factor γ is set to 0.9, 0.9, and 0.3 for data sets AWA2, CUB, and SUN, respectively. Finally, we adopt the Adam (Kingma and Ba 2015) optimizer in the experiments. The learning rate of networks AEN, PGN, and APMN all is searched in the set $\{1e^{-2}, 5e^{-3}, 1e^{-3}\}$. The learning rate of vision transformer is set to $5e^{-6}$, $5e^{-5}$ and $1e^{-6}$ for data sets AWA2, CUB and SUN, respectively. The batch size is set to 200 on all data sets. We take an image of size $3 \times 224 \times 224$ as input.

Performance Analysis

Table 2 records the experimental results of the proposed method CSAP-ZSL compared to state-of-the-art ZSL methods on data sets AWA2, CUB, and SUN. And we have the following observations

- Compared with overall embedding-based alignment methods, including generative methods and embedding methods, the proposed method CSAP-ZSL achieves higher performance on all data sets.
- Compared with part embedding-based alignment methods, including $3 \times 224 \times 224$ and $3 \times 448 \times 448$ image size as input, the proposed method CSAP-ZSL achieves higher performance on all data sets.

The proposed method CSAP-ZSL is superior to the 26 comparison methods. The possible reasons are as follows.

First, the comparison methods adopt the pixel of the feature map as the smallest unit of an image region, which is too board to capture the region corresponding to individual class semantic attributes. Instead, the proposed method CSAP-ZSL splits an image into some small regions and assigns similar regions into the same cluster. The regions of an images within the same cluster corresponds to individual class semantic attributes.

Second, the proposed method CSAP-ZSL sufficiently explores the relations between different regions of an image. Furthermore, the proposed method achieves region partition-based attribute alignment between regions of an image within the same cluster and class semantic attributes.

Third, the proposed method CSAP-ZSL achieves end-to-end training, which enables the feature extraction network to adapt to the ZSL task.

Ablation Study

We conduct ablation experiments to demonstrate the effectiveness of different components in CSAP-ZSL. Specifically, the following methods are used for comparison. (a) M_1: refers to the coarse-grained perception module and the parameters of f_e are without fine-tuning. (b) M_2: refers to the coarse-grained perception module and the parameters of f_e are fine-tuning. (c) M_3: refers to coarse-grained and fine-grained perception modules without soft region partitioning component, and the parameters of f_e are fine-tuning. (d) M_4: refers to the full model. The performance of the different models is shown in Table 3. From Table 3, we can draw the following conclusions.

First, the model M_2 achieves higher performance than the model M_1, which demonstrates that the end-to-end training can enable the pre-trained feature extraction network to adapt to the current task. Second, the performance of the

Image size	Method	End-to-End	AWA2			CUB			SUN		
			U	S	H	U	S	H	U	S	H
	Overall Embedding-based Methods										
3×224×224	QFSL (Song et al. 2018)	✓	52.1	72.8	60.7	33.3	48.1	39.4	30.9	18.5	23.1
	LDF (Li et al. 2018)	✓	9.8	87.4	17.6	26.4	81.6	39.9	–	–	–
	EBSG (Zhang, Liang, and Zhao 2022)	✓	59.3	86.7	70.4	41.8	80.7	55.1	42.6	44.1	43.3
	mVACA (Zhang et al. 2023)	✓	60.4	79.7	68.7	64.5	75.3	69.4	44.3	42.0	43.1
3×448×448	DVBE (Min et al. 2020)	✓	62.7	77.5	69.4	64.4	73.2	68.5	44.1	41.6	42.8
3×224×224	CADA-VAE (Schönfeld et al. 2019)	×	55.8	75.0	63.9	51.6	53.5	52.4	47.2	35.7	40.6
	f-VAEGAN (Xian et al. 2019b)	×	57.6	70.6	63.5	48.4	60.1	53.6	45.1	38.0	41.3
	LisGAN (Li et al. 2019)	×	–	–	–	46.5	57.9	51.6	42.9	37.8	40.2
	FREE (Chen et al. 2021a)	×	60.4	75.4	67.1	55.7	59.9	57.7	47.4	37.2	41.7
	HSVA (Chen et al. 2021b)	×	56.7	79.8	66.3	52.7	58.3	55.3	48.6	39.0	43.3
	BPL (Guan et al. 2021)	×	–	–	–	47.3	52.5	49.8	42.2	27.9	33.6
	CE-GZSL (Han et al. 2021)	×	63.1	78.6	70.0	63.9	66.8	65.3	48.8	38.6	43.1
	SCE-GZSL (Han et al. 2022)	×	64.3	77.5	70.3	66.5	68.6	67.6	45.9	41.7	43.7
	ICCE (Kong et al. 2022)	×	65.3	82.3	72.8	67.3	65.5	66.4	–	–	–
	Part Embedding-based Methods										
3×224×224	SP-AEN (Chen et al. 2018)	✓	23.3	90.9	37.1	34.7	70.6	46.6	24.9	38.6	30.3
	AREN (Xie et al. 2019)	✓	15.6	92.9	26.7	38.9	78.7	52.1	19.0	38.8	25.5
	LFGAA (Liu et al. 2019)	✓	27.0	93.4	41.9	36.2	80.9	50.0	18.5	40.0	25.3
	SGMA (Zhu et al. 2019)	✓	37.6	87.1	52.5	36.7	71.3	48.5	–	–	–
	APN (Xu et al. 2020)	✓	57.1	72.4	63.9	65.3	69.3	67.2	41.9	34.0	37.6
	DUET (Chen et al. 2023)	✓	63.7	84.7	72.7	62.9	72.8	67.5	45.7	45.8	45.8
3×448×448	VSE (Zhu, Wang, and Saligrama 2019)	✓	45.6	88.7	60.2	39.5	68.9	50.2	–	–	–
	DAZLE (Huynh and Elhamifar 2020)	×	60.3	75.7	67.1	56.7	59.6	58.1	52.3	24.3	33.2
	TransZero (Chen et al. 2022b)	×	61.3	82.3	70.2	69.3	68.3	68.6	52.6	33.4	40.8
	TransZero++ (Chen et al. 2022a)	×	64.6	82.7	72.5	67.5	73.6	70.4	48.6	37.8	42.5
	MSDN (Chen et al. 2022d)	×	62.0	74.5	67.7	68.7	67.5	68.1	52.2	34.2	41.3
	GNDAN (Chen et al. 2022c)	×	60.2	80.8	69.0	69.2	69.6	69.4	50.0	34.7	41.0
3×224×224	**CSAP-ZSL**	✓	66.8	84.4	**74.6**	68.1	73.2	**70.6**	61.3	45.5	**52.2**

Table 2: Experimental results (%) compared to state-of-the-art ZSL methods on data sets AWA2, CUB, and SUN. The best results are marked in bold and the second-best results are marked in underlined. The symbol "–" indicates no results in original literature.

Model	AWA2			CUB			SUN		
	U	S	H	U	S	H	U	S	H
M_1	51.2	87.7	64.7	54.8	65.7	59.8	56.4	38.7	45.9
M_2	58.5	81.7	68.2	63.4	66.9	65.1	46.9	51.0	48.9
M_3	59.7	87.9	71.1	61.8	74.8	67.7	49.2	50.9	50.0
M_4	66.8	84.4	74.6	68.1	73.2	70.6	61.3	45.5	52.2

Table 3: Experimental results (%) of ablation analysis of the proposed method on data sets AWA2, CUB, and SUN.

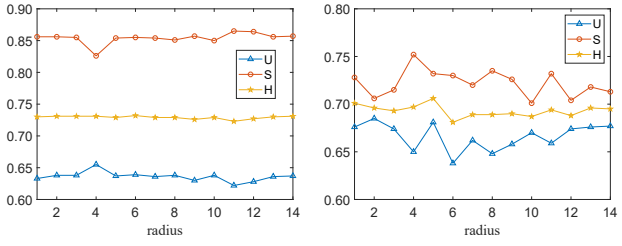


Figure 3: The variation tendency of the U, S, and H as the radius increases.

model M_3 is superior to the model M_2, because the fine-grained perception module can provide more discriminative information for the model. Third, the full model M_4 achieves the highest performance, which verifies the effectiveness of soft region partitioning and region partition-based attribute alignment components.

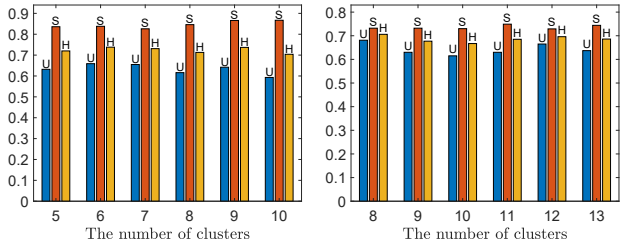
Parameters Study

This section shows the experimental results on data sets AWA2 and CUB.

The radius of graph Fig. 3 shows the variation tendency of the U, S, and H of the proposed method CSAP-ZSL under graphs of regions with different radius settings. In the experiment, the radius varies within $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$. When the radius is set to 6 and 5, the H of the proposed method CSAP-ZSL is best on data sets AWA2 and CUB, respectively. And the overall performance trend is relatively gentle under different radius settings on all data sets.

The number of clusters Fig. 4 shows the variation tendency of the U, S, and H of the proposed method CSAP-ZSL with different numbers of clusters in graph cut. In the experiment, the number of the clusters varies within $\{5, 6, 7, 8, 9, 10\}$ for data sets AWA2 and within

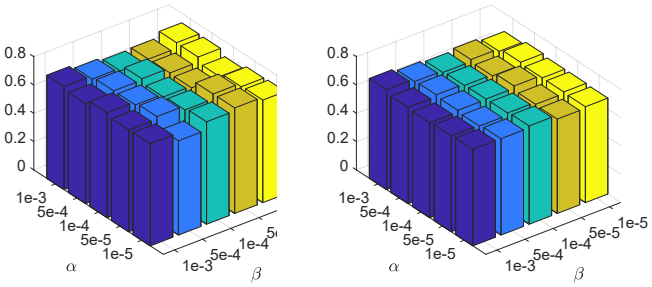
$\{8, 9, 10, 11, 12, 13\}$ for data set CUB. When the number of clusters is set to 6 and 8, the H of the proposed method CSAP-ZSL is best on data sets AWA2 and CUB, respectively. The overall trend of performance is relatively gentle under different numbers of clusters on all data sets.



(a) AWA2 (b) CUB

Figure 4: The variation tendency of the U, S, and H with different numbers of clusters in graph cut.

The trade-off parameters Fig. 5 shows the variation tendency of the H of the proposed method CSAP-ZSL with different trade-off parameters α and β. From Fig. 5, we can observe that the H of the proposed method CSAP-ZSL is not sensitive to hyper-parameters α and β variations on data sets AWA2 and CUB. Moreover, we obtain competitive performance of the proposed method CSAP-ZSL within a limited range compared to the state-of-the-art methods.



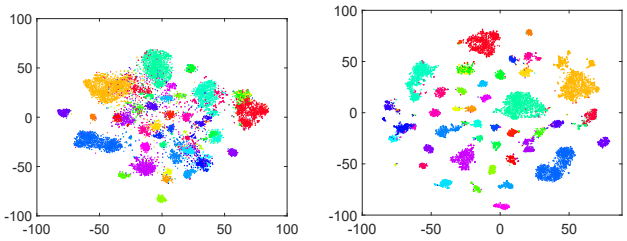
(a) AWA2 (b) CUB

Figure 5: The variation tendency of the H with different trade-off parameters α and β.

Visualization Study

Taking data set AWA2 as an example, we visualize the T-SNE (van der Maaten and Hinton 2008) feature embeddings extracted by the pre-trained ResNet101 and fine-tuning ViT-base in the proposed method CSAP-ZSL, respectively. As shown in Fig. 6, the feature embeddings of the proposed method are more discriminative than the pre-trained ResNet101.

In addition, because the class semantic attributes on the data set CUB can be directly reflected in images, we show the head regions of images obtained by the proposed method on the data set CUB. As shown in Fig. 7, each image is divided into 8 clusters by broken line. The proposed method



(a) (b)

Figure 6: The T-SNE visualization. (a) pre-trained ResNet101. (b) fine-tuning encoder (ViT-base) in the proposed method CSAP-ZSL.

discovers the regions corresponding to the head by soft region partitioning. Although the heads are located in different positions in different images, the cluster of regions corresponding to the head get the maximum response.

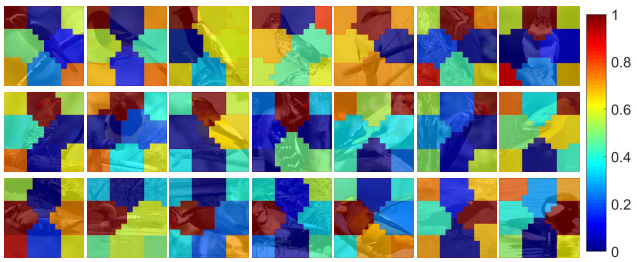


Figure 7: The exemplars for the region partitioning result of the image in the proposed method CSAP-ZSL on data set CUB. Different colors represent different levels of response to attribute "head".

Conclusion

This paper proposes a class semantic attribute perception guided ZSL method, which includes fine-grained and coarse-grained perception modules. The fine-grained perception module splits an image into some regions and then explores the relations between regions of an image by soft region partitioning. Furthermore, this module achieves region partition-based attribute alignment between the regions of an image within the same cluster and class semantic attributes. The coarse-grained perception module achieve coarse-grained perception of class semantic attributes across the entire image through contrastive semantic learning. These two modules are united into a network, achieving an end-to-end model. Finally, experiments on ZSL data sets verify the effectiveness of the proposed method.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020AAA0106102) and the National Natural Science Foundation of China (Nos. 62432006, 62376141).

References

Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; and Chang, S. 2018. Zero-Shot Visual Recognition Using Semantics-Preserving Adversarial Embedding Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, 1043–1052.

Chen, S.; Hong, Z.; Hou, W.; Xie, G.-S.; Song, Y.; Zhao, J.; You, X.; Yan, S.; and Shao, L. 2022a. TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–17.

Chen, S.; Hong, Z.; Liu, Y.; Xie, G.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; and You, X. 2022b. TransZero: Attribute-Guided Transformer for Zero-Shot Learning. In *AAAI Conference on Artificial Intelligence, Virtual Event*, 330–338.

Chen, S.; Hong, Z.; Xie, G.; Peng, Q.; You, X.; Ding, W.; and Shao, L. 2022c. GNDAN: Graph Navigated Dual Attention Network for Zero-Shot Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Chen, S.; Hong, Z.; Xie, G.; Yang, W.; Peng, Q.; Wang, K.; Zhao, J.; and You, X. 2022d. MSDN: Mutually Semantic Distillation Network for Zero-Shot Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 7602–7611.

Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; and Shao, L. 2021a. FREE: feature refinement for generalized zero-shot learning. In *IEEE International Conference on Computer Vision, Montreal, QC, Canada*, 122–131.

Chen, S.; Xie, G.; Peng, Q.; Liu, Y.; Sun, B.; Li, H.; You, X.; and Shao, L. 2021b. HSVA: hierarchical semantic-visual adaptation for zero-shot learning. In *Advances in Neural Information Processing Systems, Virtual Event*, 16622–16634.

Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Zhang, W.; Fang, Y.; Pan, J. Z.; and Chen, H. 2023. DUET: Cross-Modal Semantic Grounding for Contrastive Zero-Shot Learning. In *AAAI Conference on Artificial Intelligence, Washington, DC, USA*, 405–413.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations, Virtual Event, Austria*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, United States*, 2121–2129.

Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2015. Transductive Multi-View Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11): 2332–2345.

Fu, Z.; Xiang, T.; Kodirov, E.; and Gong, S. 2018. Zero-Shot Learning on Semantic Class Prototype Graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8): 2009–2022.

Guan, J.; Lu, Z.; Xiang, T.; Li, A.; Zhao, A.; and Wen, J. 2021. Zero and few shot learning with semantic feature synthesis and competitive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2510–2523.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021. Contrastive Embedding for Generalized Zero-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, virtual*, 2371–2381.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2022. Semantic Contrastive Embedding for Generalized Zero-Shot Learning. *International Journal of Computer Vision*, 130(11): 2606–2622.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 15979–15988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA*, 770–778.

Huynh, D.; and Elhamifar, E. 2020. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA*, 4482–4492.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, San Diego, CA, USA*.

Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic Autoencoder for Zero-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, 4447–4456.

Kong, X.; Gao, Z.; Li, X.; Hong, M.; Liu, J.; Wang, C.; Xie, Y.; and Qu, Y. 2022. En-Compactness: Self-Distillation Embedding & Contrastive Generation for Generalized Zero-Shot Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 9296–9305.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA*, 951–958.

Larochelle, H.; Erhan, D.; and Bengio, Y. 2008. Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence, Chicago, Illinois, USA*, 646–651.

Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; and Huang, Z. 2019. Leveraging the invariant side of generative zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, 7402–7411.

Li, J.; Lan, X.; Long, Y.; Liu, Y.; Chen, X.; Shao, L.; and Zheng, N. 2020. A Joint Label Space for Generalized Zero-Shot Classification. *IEEE Transactions on Image Processing*, 29: 5817–5831.

Li, Y.; Zhang, J.; Zhang, J.; and Huang, K. 2018. Discriminative Learning of Latent Features for Zero-Shot Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, 7463–7471.

Liu, Y.; Guo, J.; Cai, D.; and He, X. 2019. Attribute Attention for Semantic Disambiguation in Zero-Shot Learning. In *IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South)*, 6697–6706.

Min, S.; Yao, H.; Xie, H.; Wang, C.; Zha, Z.; and Zhang, Y. 2020. Domain-Aware Visual Bias Eliminating for Generalized Zero-Shot Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA*, 12661–12670.

Narayan, S.; Gupta, A.; Khan, S. H.; Khan, F. S.; Shao, L.; and Shah, M. 2021. Discriminative Region-based Multi-Label Zero-Shot Learning. In *IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada*, 8711–8720.

Patterson, G.; and Hays, J. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA*, 2751–2758.

Reed, S. E.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA*, 49–58.

Romera-Paredes, B.; and Torr, P. H. S. 2015. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning, Lille, France*, volume 37, 2152–2161.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252.

Schönfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero- and few-shot learning via aligned variational autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, 8247–8255.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations, San Diego, CA, USA*.

Song, J.; Shen, C.; Yang, Y.; Liu, Y.; and Song, M. 2018. Transductive Unbiased Embedding for Zero-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, 1024–1033.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA*, 1–9.

Tang, C.; He, Z.; Li, Y.; and Lv, J. 2022. Zero-Shot Learning via Structure-Aligned Generative Adversarial Network. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11): 6749–6762.

van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD birds 200.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2019a. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9): 2251–2265.

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019b. F-VAEGAN-D2: A feature generating framework for any-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, 10275–10284.

Xie, G.; Liu, L.; Jin, X.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.; and Shao, L. 2019. Attentive Region Embedding Network for Zero-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, 9384–9393.

Xu, B.; Zeng, Z.; Lian, C.; and Ding, Z. 2022a. Generative Mixup Networks for Zero-Shot Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute prototype network for zero-shot learning. In *Advances in Neural Information Processing Systems, Virtual Event*, 21969–21980.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2022b. VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 9306–9315.

Zhang, C.; Liang, C.; and Zhao, Y. 2022. Exemplar-Based, Semantic Guided Zero-Shot Visual Recognition. *IEEE Transactions on Image Processing*, 31: 3056–3065.

Zhang, H.; Tian, L.; Wang, Z.; Xu, Y.; Cheng, P.; Bai, K.; and Chen, B. 2023. Multiscale Visual-Attribute Co-Attention for Zero-Shot Image Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9): 6003–6014.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a Deep Embedding Model for Zero-Shot Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, 3010–3019.

Zhu, P.; Wang, H.; and Saligrama, V. 2019. Generalized Zero-Shot Recognition Based on Visually Semantic Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, 2995–3003.

Zhu, Y.; Xie, J.; Tang, Z.; Peng, X.; and Elgammal, A. 2019. Semantic-Guided Multi-Attention Localization for Zero-Shot Learning. In *Advances in Neural Information Processing Systems, Vancouver, BC, Canada*, 14917–14927.